Quantitative Analysis

ECO375 Summer

Professor Yuanyuan Wan

Shih-Chieh Lee

Due: June 25, 2022

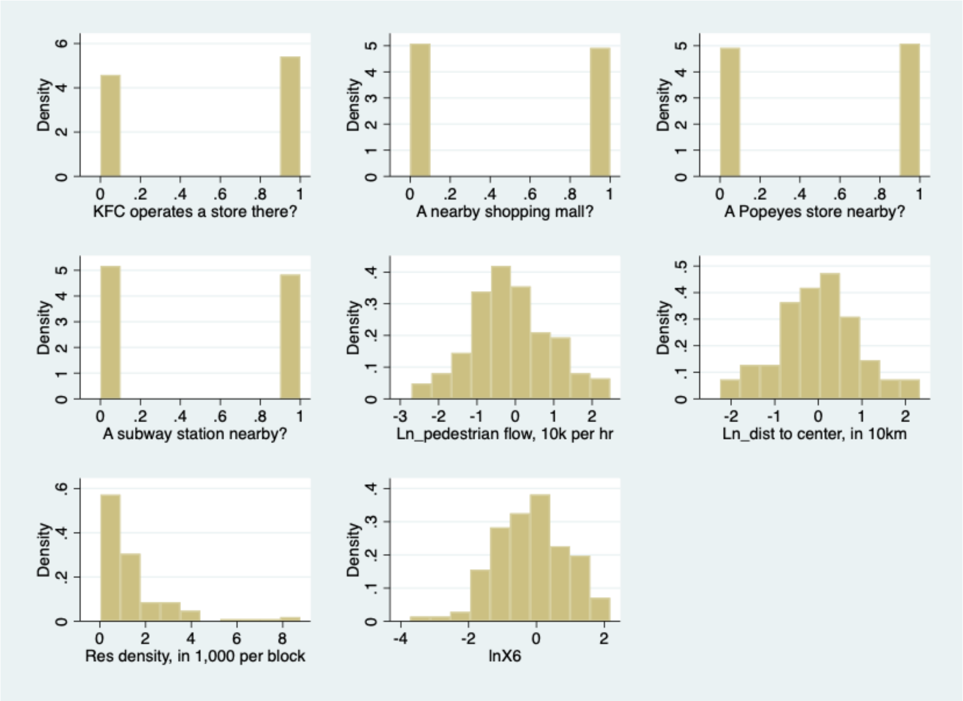
Background and Aims

This report is inspired by MacDonald (referred to as “the client” in the following)’s plan to open a new store at the Eaton Center. In order to maximize the client’s profits, the purpose of this report is to analyze the probability for KFC, one of the client’s major competitors in the fast food industry, to open a new store at the Eaton Center.

Descriptive Statistics

This report conducts data analysis based on the provided dataset on 120 locations in the Greater Toronto Area. The dataset contains 7 variables with no missing observation of any kind. Variables y, x1, x2, and x3 are binary variables, where y represents whether KFC operates a store in the given location (denoted as 1) or not (denoted as 0); similarly, variable x1, x2, x3 indicates if the given location has a shopping mall, a Popeyes Fried Chicken nearby, or a metro station nearby or not respectively, where 1 refers to affirmation and 0 refers to negation for its corresponding statement. According to Stata output, all binary variables in the dataset have a mean and a standard deviation close to 0.5, implying they are close to evenly distributed.

Variables x4, x5, and x6 are continuous variables that potentially affect KFC’s decision to operate a store. x4 specifies the natural logarithm of the pedestrian flow at the nearest major intersection, where the pedestrian flow is 10,000 people per hour. x5 shows the level of convenience from a logistic perspective, measuring the natural logarithm of the distance between the given location and the nearest KFC distribution center (distance measured in 10 kilometers). x6 reflects the residency density of the given location, which is measured in 1,000 people per block. According to density histograms generated by Stata (Figure 1), the distributions of x4 and x5 demonstrates some proximity to a normal distribution whereas the distribution of x6 is extremely skewed towards right, suggesting most locations in the dataset are in blocks with a residency density lower than 1,000 people.

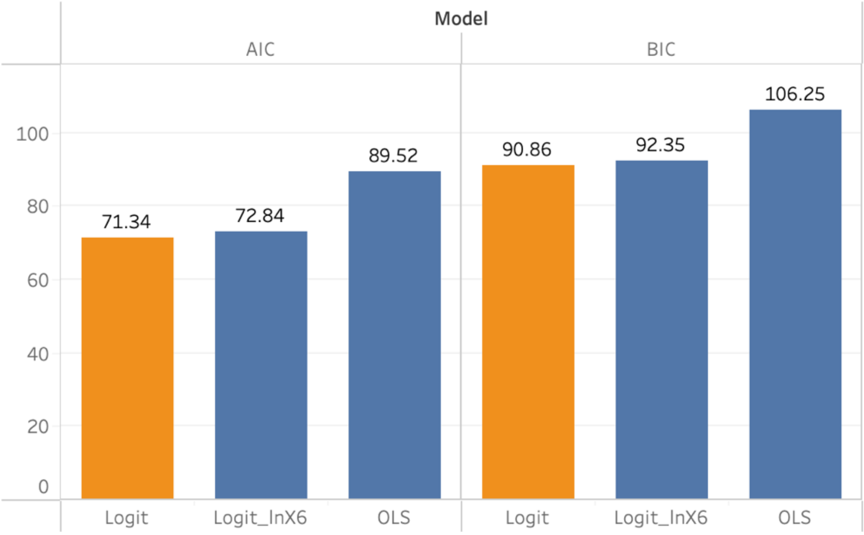


[Figure 1: Histograms for y, x1~x6, lnX6, respectively]

Methodology & Assumptions

Given the binary nature of our dependent variable, y, the model this report adopts primarily is the binary response model using the logit regression. Thus, we use stepwise function to generate the more predictive model from the given dataset, and model $P(y=1|x)=\Lambda(\beta 0+\beta 1x1+\beta 2x2+\beta 3x3+\beta 4x4+\beta 5x5+\beta 6x6)$, where $\beta 0$ ~ $\beta 6$ represents the estimate coefficients for each of their corresponding regressor variables, appears to be the best possible model drawn from our current dataset, according to Stata.

Despite being proven less effective in the case of the binary dependent variable by theory, we still choose to establish an OLS model using Stata stepwise function and compare it to the previous probit model. Also, the skewness of x6 distribution is worth noticing, hence we also create a new variable lnX6 obtained from the natural logarithm of x6 value, and run a new stepwise logit regression replacing x6 with lnX6. Both of the new models, however, are inferior according to results from Stata information criteria. Qualitatively speaking, an OLS model, if without interaction terms, only generates a constant estimate coefficient whereas the binary response model provides more empirically-robust partial effects of x on the dependent variable. Therefore, this report selects the logit model where y is the dependent variable, and x1, x2, x3, x4, x5, and x6 are independent variables.



[Figure 2: Model Selection]

Before using the selected model, its implied assumptions should also be introduced and discussed. First, the regressor variables must be independent of the residual. Second, no perfect multicollinearity should occur among the selected regressor variables. We will further discuss these assumptions later.

Data Analysis & Prediction

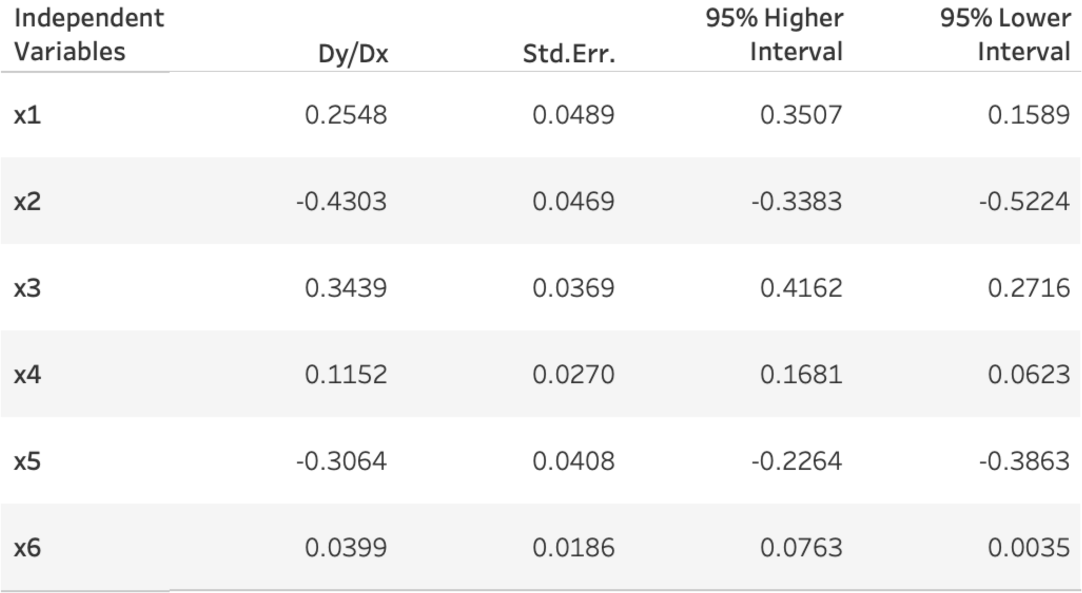
Applying the selected model to the dataset, we obtain the following result:

$$P(y=1|x) = \Lambda(-1.565264 + 3.415127x_1 - 5.768404x_2 + 4.609453x_3 + 1.544106x_4 - 4.106603x_5 + 0.5345794x_6)$$

where each of the individual regressors is significant at 5% level, and the model overall is significant at any conventional significance level with substantial goodness of fitness (pseudo-R2 = 0.6536). According to our model prediction, it is almost certain that KFC would operate a new store at the Eaton Center (Pr(y=1|x) = 0.999996). Empirically speaking, It is also natural to set up a fast food store in one of the most crowded regions in the GTA area, hence the prediction is not counterintuitive in any common sense. However, there are more interesting findings to be further addressed within our data analysis.

The partial effects of each independent variable on the outcome provide us an insight into KFC's consideration behind this decision. Among those factors represented in continuous variables (x4, x5, x6), it is the location's distance to the nearest KFC distribution center that affects KFC's decision the most. On average, locations that are 10% farther from the distribution center are 3.06 percent points less likely for KFC to operate a new store there, assuming all other 5 factors are equal. The pedestrian flow, on the other hand, influences KFC's decision positively to a large extent. Locations with 10% more pedestrian flow are about 1.2 percentage points less likely to have a new KFC store. However, it seems like KFC is relatively less concerned about the resident density of a location when they reach their decision – on average, locations with a stunning 1,000 increase in the block residency density barely boost the chance to have a new KFC store by 0.03 percentage points. Hence, we conclude that the residency density of a location is of little concern when KFC chooses where to operate a new store.

We also conduct a marginal analysis on the partial effects of 3 binary dummy variables (x1 x2 x3) on the probability of having a new KFC store in a location. The results show that all 3 indicators, namely whether a shopping mall, a Popeyes Fried Chicken, or a metro station is located nearby have little influence on KFC's decision on having a new store. When there is a Popeyes Fried Chicken nearby, KFC is only 0.4303 percent points less likely to operate a store compared to a location without a Popeyes Fried Chicken, assuming all other 5 factors are the same. Hence, we conclude that KFC mainly consults factors like the location's distance to its distribution center and the pedestrian flow. It moderately considers the residency population but remains aloof toward factors like the presence of its competitor's store, metro station, and shopping mall.

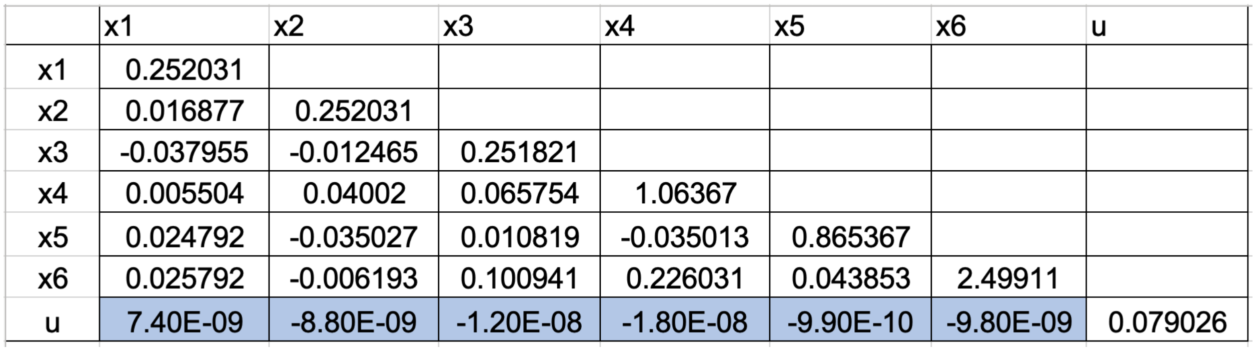


Independent Variables	Dy/Dx	Std.Err.	95% Higher Interval	95% Lower Interval
x1	0.2548	0.0489	0.3507	0.1589
x2	-0.4303	0.0469	-0.3383	-0.5224
x3	0.3439	0.0369	0.4162	0.2716
x4	0.1152	0.0270	0.1681	0.0623
x5	-0.3064	0.0408	-0.2264	-0.3863
x6	0.0399	0.0186	0.0763	0.0035

[Figure 3: The average partial effects of each regressors]

Discussion

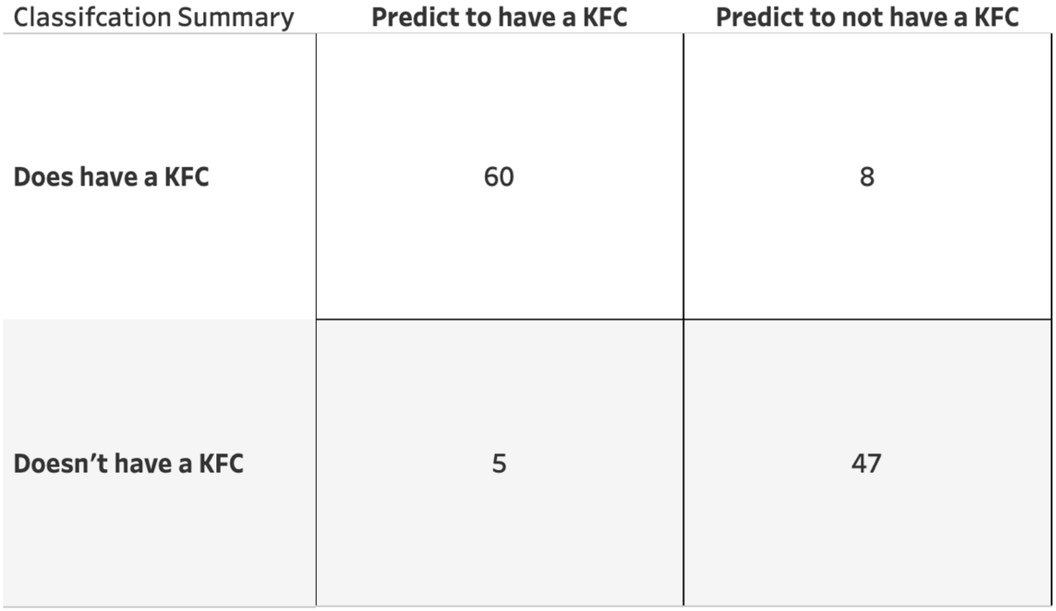
A few assumptions are made when we were establishing our model, and this requires further justification. First, we assume no perfect multicollinearity among any regressors, and this can be simply justified qualitatively. Speaking from the common sense, none of these variables could be possibly measured in any form of linear combination of other variables. Second, we assume all regressors variables must be independent from the residual. To justify this assumption, we use Stata to calculate our the residual between the predicted probability and the actual probability of having a KFC store, and construct a variance-covariance table between the residual and all of regressors in our model, which is reported below as Figure 4. The insignificant covariance between any regressor and the residual demonstrates the exogeneity of our dependent variables, thus the assumption can be justified.



	x1	x2	x3	x4	x5	x6	u
x1	0.252031						
x2	0.016877	0.252031					
x3	-0.037955	-0.012465	0.251821				
x4	0.005504	0.04002	0.065754	1.06367			
x5	0.024792	-0.035027	0.010819	-0.035013	0.865367		
x6	0.025792	-0.006193	0.100941	0.226031	0.043853	2.49911	
u	7.40E-09	-8.80E-09	-1.20E-08	-1.80E-08	-9.90E-10	-9.80E-09	0.079026

[Figure 5: The variance-covariance table among u and x']

The limit of our model should also be discussed. First, according to Stata specification test, predictions generated from our model tends to be more accurate when the location has a KFC store in reality. Despite successfully predicting more than 89% of observations using our model, the classification reveals a higher rate of misprediction for no KFC cases, demonstrating our model's bias to overpredict positive outcome. Therefore, this model requires further revisions by including more observations.



Classifcation Summary	**Predict to have a KFC**	**Predict to not have a KFC**
Does have a KFC	60	8
Doesn't have a KFC	5	47

[Figure 5: Results from Stata Classification Summary]

Also, our model is established based on this limited dataset. However, the Eaton is a location that is 15 kilometers far from the nearest KFC and has a residency population of 4500, which are beyond any observations we have in the given dataset. The credibility of the predicted result, hence, should not be exaggerated to an extent given this situation.

Last but not the least, due to the limit of the available dataset, this report is unable to put it into the modeling process and hence the reliability of the following prediction may be dramatically undermined.